

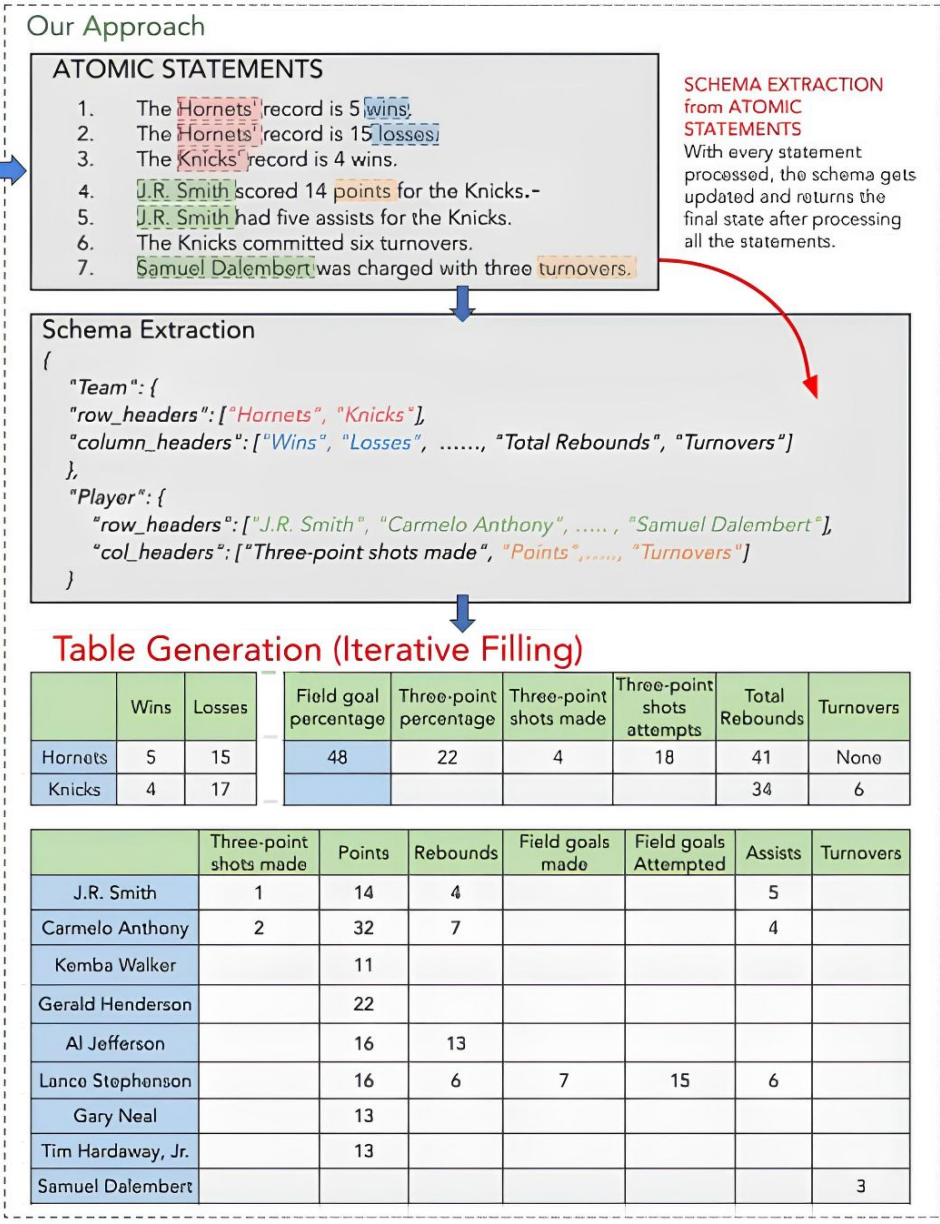
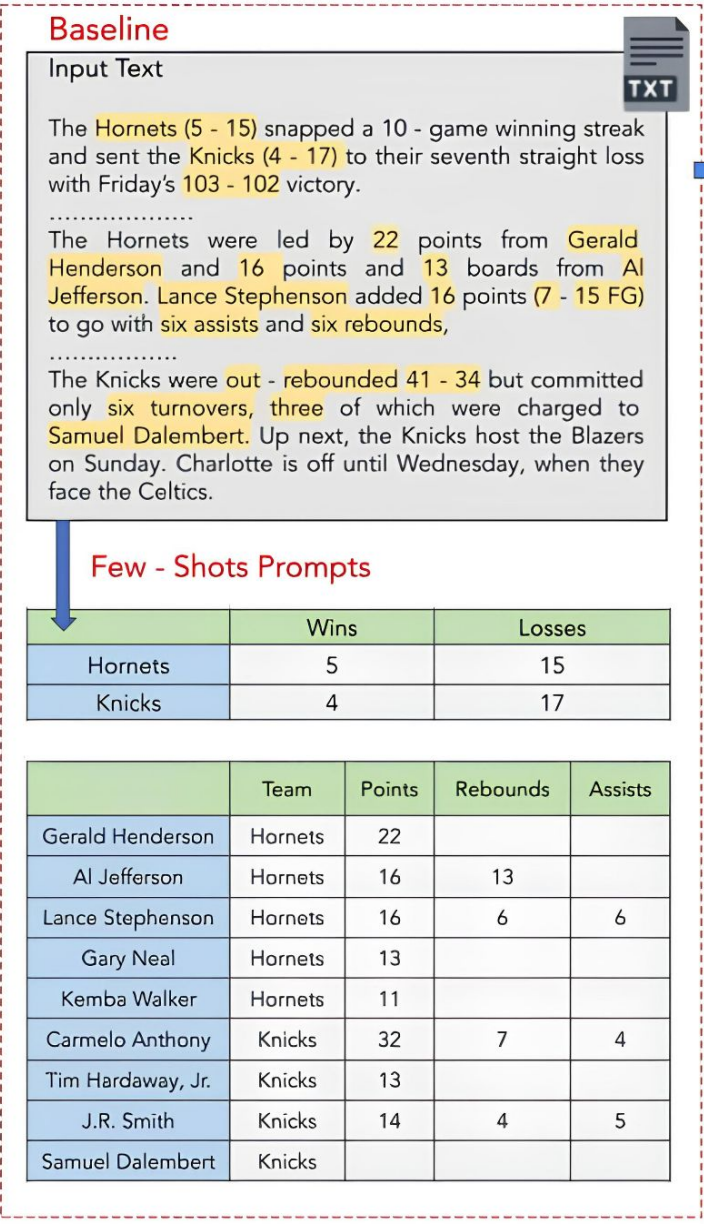
## Motivation

- Limitation of Predefined Schemas:** Current methods require a predefined table structure (the columns), which is impractical for real-world text where the structure is unknown.
- Inadequacy in Handling Complexity:** Existing methods are too simple and fail to extract complex relationships, infer data, or perform numerical aggregation from dense narratives.

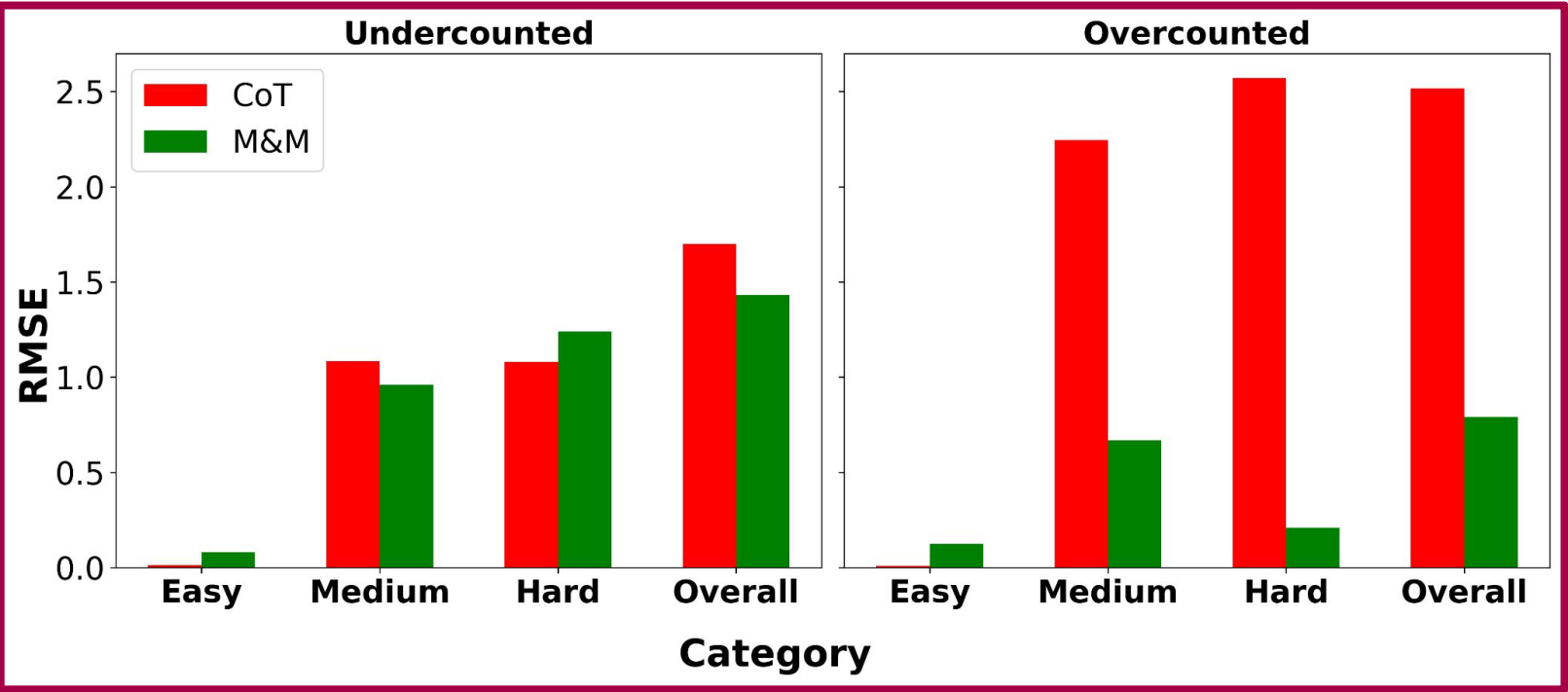
## Our Contribution

- Introduce a **generalized notion of Structured Summarization**, evaluating LLMs' capabilities on planning table schemas, summarize information under a given instruction exhaustively.
- Generalizable framework**, applicable to any Instruction-Driven Tabular Summarization task extending beyond simple extraction tasks
- Manually corrected version of the Rotowire Benchmark, a popular text-to-table benchmark.
- Comprehensive evaluation of LLMs on a **diverse suite of metrics encompassing** both table quality and information coverage.

## Methodology

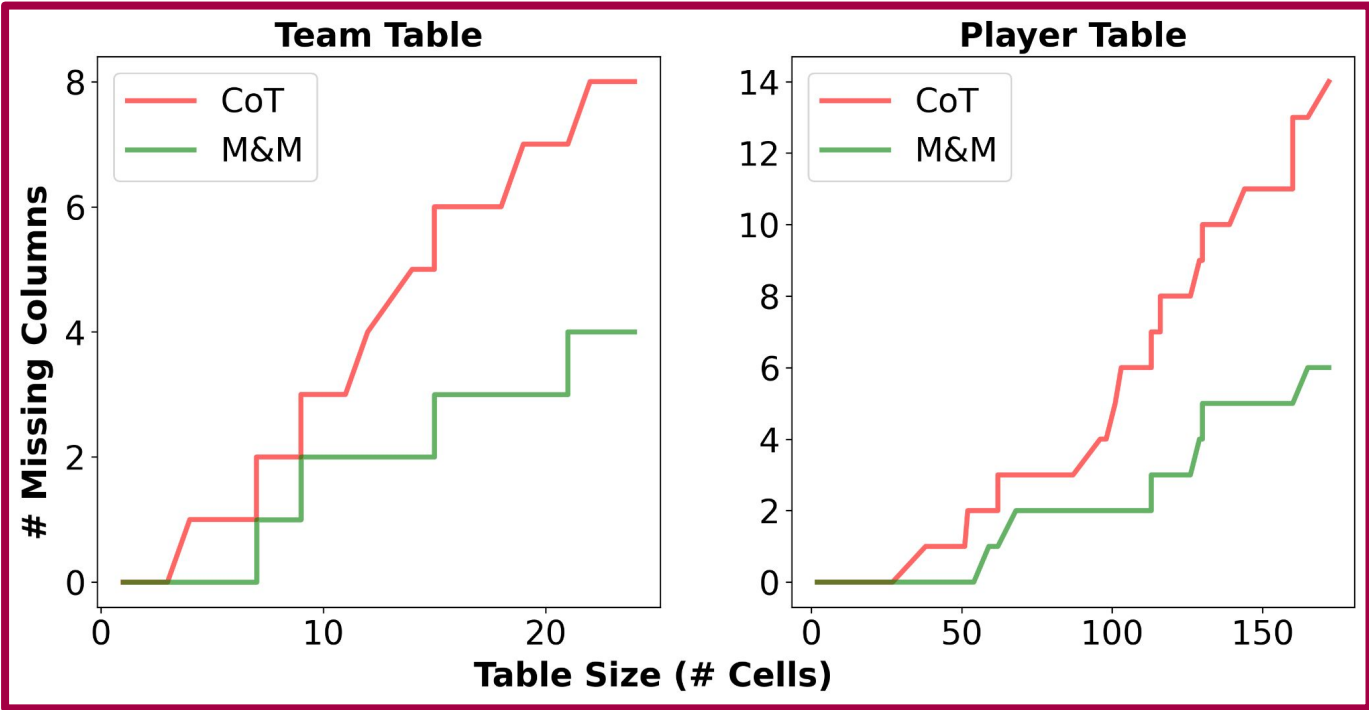


## Analysis



- CoT:** shows high overcounting errors i.e overcounting events .
- Map&Make:** shows stable performance across different difficulty levels.

- CoT:** Scales poorly, with missing columns increasing sharply as tables grow.
- Map&Make:** Scales well, maintaining a stable schema with fewer missing columns.



## Results

### Rotowire

- Completeness:** improves completeness scores increasing by up to 42%.
- Correctness:** Lower Correctness scores are misleading; they result from M&M capturing valid, non-statistical data missing from the ground truth.
- Validation:** AutoQA scores confirm M&M's higher actual accuracy.

Method	TabEval						Auto-QA
	Correctness		Completeness		Overall		Accuracy
	Team	Player	Team	Player	Team	Player	
GPT-4o							
CoT	80.23	86.74	56.26	57.74	64.34	65.57	41.42
T <sup>3</sup> - D	80.24	<b>87.56</b>	56.12	64.34	64.09	70.84	38.08
T <sup>3</sup>	<b>81.65</b>	85.39	46.43	60.66	56.05	66.61	39.46
M&M - U	66.28	86.16	<b>77.65</b>	<b>88.15</b>	<b>69.56</b>	<b>85.85</b>	69.51
M&M - 3S	48.78	74.22	61.43	87.34	51.01	78.56	<b>80.38</b>
Gemini-2.0							
CoT	<b>82.91</b>	<b>91.51</b>	63.50	75.89	69.72	81.06	60.52
T <sup>3</sup> - D	81.86	91.08	65.18	76.27	70.01	80.92	47.94
T <sup>3</sup>	75.70	87.22	61.75	82.07	64.93	82.68	56.58
M&M - U	75.27	91.17	77.47	92.92	<b>74.80</b>	<b>91.36</b>	64.88
M&M - 3S	65.89	83.54	<b>79.49</b>	<b>92.43</b>	69.95	86.71	<b>71.21</b>

Method	AutoQA-Score
Chain of Thoughts	63.68
Text-Tuple-Table	67.07
Map & Make (3-step)	<b>76.40</b>

### Wiki40B

- M&M:** Shows higher information coverage
- Dynamically adapts to multi-table generation (**min:1, max:13, avg: 6.2**)
- Generalises to **open-domain text**

Method	Easy		Medium		Hard		Average	
	RMSE	ER	RMSE	ER	RMSE	ER	RMSE	ER
<b>GPT-4o</b>								
CoT	0.05	2.62	2.05	55.30	1.73	76.86	1.61	45.34
T <sup>3</sup> D	0.08	<b>0.43</b>	1.46	46.62	2.73	84.73	1.75	44.91
T <sup>3</sup>	0.23	16.5	1.48	34.16	2.39	51.91	1.76	35.62
M&M - U	<b>0.03</b>	1.85	0.81	31.06	1.96	68.10	1.19	32.94
M&M - 3S	0.07	3.44	<b>0.63</b>	<b>25.48</b>	<b>0.93</b>	<b>32.34</b>	<b>0.71</b>	<b>21.77</b>
<b>Gemini-2.0</b>								
CoT	0.01	0.63	1.69	58.31	2.23	82.80	1.68	48.32
T <sup>3</sup> D	<b>0.01</b>	<b>0.54</b>	1.21	49.63	2.34	85.12	1.49	46.44
T <sup>3</sup>	0.19	9.26	0.93	26.32	2.51	54.72	1.59	29.10
M&M - U	0.04	1.45	0.89	36.64	1.98	69.93	1.21	34.65
M&M - 3S	0.10	4.58	<b>0.60</b>	<b>23.72</b>	<b>0.89</b>	<b>33.13</b>	<b>0.70</b>	<b>21.37</b>

### Livesum

- M&M:** Reduces error rates by up to 55% and RMSE by 57% in one-shot settings.
- CoT:** Outperforms T<sup>3</sup>-D by using code-generation to integrate and count data.
- T<sup>3</sup>-D:** Shows poor generalization in schema-agnostic settings.

Corresponding Email:  
vgupt140@asu.edu  
(Vivek Gupta)

Website:  
<https://map-make.github.io/>



Scan QR for the paper